A Minimalistic Approach to Appearance based Visual SLAM

Henrik Andreasson, Tom Duckett, and Achim J. Lilienthal

***Abstract*—This paper presents a vision-based approach to SLAM in indoor / outdoor environments with minimalistic sensing and computational requirements. The approach is based on a graph representation of robot poses, using a relaxation algorithm to obtain a globally consistent map. Each link corresponds to a relative measurement of the spatial relation between the two nodes it connects. The links describe the likelihood distribution of the relative pose as a Gaussian distribution. To estimate the covariance matrix for links obtained from an omni-directional vision sensor, a novel method is introduced based on the relative similarity of neighbouring images. This new method does not require determining distances to image features using multiple view geometry, for example. Combined indoor and outdoor experiments demonstrate that the approach can handle qualitatively different environments (without modification of the parameters), that it can cope with violations of the "flat floor assumption" to some degree, and that it scales well with increasing size of the environment, producing topologically correct and geometrically accurate maps at low computational cost. Further experiments demonstrate that the approach is also suitable for combining multiple overlapping maps, e.g. for solving the multi-robot SLAM problem with unknown initial poses.**

***Index Terms*—SLAM, Omnidirectional Vision**

I. Introduction

This paper presents a new vision-based approach to the problem of simultaneous localization and mapping (SLAM). Especially compared to SLAM approaches using a 2-d laser scanner, the rich information provided by a vision-based approach about a substantial part of the environment allows for dealing with high levels of occlusion [1] and enables solutions that do not rely strictly on a flat floor assumption. Cameras can also offer a longer range and are therefore advantageous in environments that contain large open spaces.

The proposed method is called "Mini-SLAM" since it is minimalistic in several ways. On the hardware side, it relies solely on odometry and an omni-directional camera as the external source of information. This allows for less expensive systems compared to methods that use 2-d or 3-d laser scanners. Please note that the robot used for the experiments was also equipped with a 2-d laser scanner. This laser scanner, however, was not used in the SLAM algorithm but only to visualize the consistency of the created maps.

Apart from the frugal hardware requirements, the method is also minimalistic in its computational demands. Map estimation is performed on-line by a linear time SLAM algorithm on an efficient graph representation. The main difference to other vision-based SLAM approaches is that there is no estimate of the positions of a set of landmarks involved, enabling the algorithm to scale up better with the size of the environment. Instead, a measure of image similarity is used to estimate the relative pose between corresponding images ("visual relations") and the uncertainty of this estimate. Given these "visual relations" and relative pose estimates between consecutive images obtained from the odometry of the robot ("odometry relations"), the Multilevel Relaxation algorithm [2] is then used to determine the maximum likelihood estimate of all image poses. The relations are expressed as a relative pose estimate and the corresponding covariance. A key insight is that the estimate of the relative pose in the "visual relations" does not need to be very accurate as long as the corresponding covariance is modeled appropriately. This is because the relative pose is only used as an initial estimate that the Multilevel Relaxation algorithm can adjust according to the covariance of the relation. Therefore, even with fairly imprecise initial estimates of the relative poses it is possible to build geometrically accurate maps using the geometric information in the covariance of the relative pose estimates. Mini-SLAM was found to produce consistent maps in various environments, including, for example, a data set of an environment containing indoor and outdoor passages (path length of 1.4 km) and an indoor data set covering five floor levels of a department building.

Further to our previously published work [3], we extended the Mini-SLAM approach to the multi-robot SLAM problem, demonstrating its ability to combine multiple overlapping maps with unknown initial poses. We also provide an evaluation of the robustness of the suggested approach with respect to poor odometry or a less reliable measure of visual similarity.

A. Related Work

Using a camera as the external source of information in SLAM has received increasing attention during the past years. Many approaches extract landmarks using local features in the images and track the positions of these landmarks. As the feature descriptor, Lowe's scale invariant feature transform (SIFT) [4] has been used widely [5], [6]. An initial estimate of the relative pose change is often obtained from odometry [6], [7], [8], or where multiple cameras are available as in [9], [10], multiple view geometry can be applied to obtain depth estimates of the extracted features. To update and maintain visual landmarks, Extended Kalman Filters (EKF) [7], [11] and Rao-Blackwellised Particle Filters (RBPF) [6], [9] have been used. In the visual SLAM method proposed in [11] particle filters were utilised to obtain the depth of landmarks

H. Andreasson and A. J. Lilienthal are with the Department of Technology, Örebro University, Sweden e-mail: {henrik.andreasson, achim.lilienthal}@tech.oru.se.

T. Duckett is with the Department of Computer Science, University of Lincoln, UK e-mail: tduckett@lincoln.ac.uk

while the landmark positions were updated with an EKF. Initial landmark positions had to be provided by the user. A similar approach described in [8] applies a converse methodology. The landmark positions were estimated with a Kalman filter (KF) and a particle filter was used to estimate the path.

Due to their suitability for addressing the correspondence problem, vision-based systems have been applied as an addition to laser scanning based SLAM approaches for detecting loop closure. The principle has been applied to SLAM systems based on a 2D laser scanner [12] and a 3D laser scanner [13].

In the approach proposed in this paper, the SLAM optimization problem is solved at the graph-level with the Multilevel Relaxation (MLR) method of Frese and Duckett [2]. This method could be replaced by alternative graph based SLAM methods, for example, the online method proposed by Grisetti et al. [14] based on the stochastic gradient descent method proposed by Olson et al. [15].

The rest of this paper is structured as follows. Section II describes the proposed SLAM approach. Then the experimental set-up is detailed and the results are presented in Section III. The paper ends with conclusions and suggestions for future work (Section IV).

II. MINI-SLAM

A. Multi-Level Relaxation

The SLAM optimization problem is solved at the graph-level with the Multilevel Relaxation (MLR) method of Frese and Duckett [2]. A map is represented as a set of nodes connected in a graph structure. An example is shown in Fig. 1. Each node corresponds to the robot pose at a particular time and each link to a relative measurement of the spatial relation between the two nodes it connects. A node is created for each omni-image in this work and the terms node and frame are used interchangeably in this paper.

The MLR algorithm can be briefly explained as follows. The input is a set $\mathcal{R}$ of $m = |\mathcal{R}|$ relations on n planar frames (i.e., a two-dimensional representation is used). Each relation $r \in \mathcal{R}$ describes the likelihood distribution of the pose of frame a relative to frame b. Relations are modeled as a Gaussian distribution with mean μ^r and covariance C^r. The output of the MLR algorithm is the maximum likelihood estimation vector $\hat{x}$ for the poses of all the frames. Thus, a globally consistent set of Cartesian coordinates is obtained for the nodes of the graph based on local (relative) and inconsistent (noisy) measurements, by maximizing the total likelihood of all measurements.

B. Odometry Relations

The Mini-SLAM approach is based on two principles. First, that odometry is sufficiently accurate if the distance traveled is short. Second, that by using visual matching, correspondence between robot poses can be detected reliably even though the search region around the current pose estimate is large. Accordingly, two different types of relations are created in the MLR graph, based on odometry r_o and relations based on visual similarity r_v.

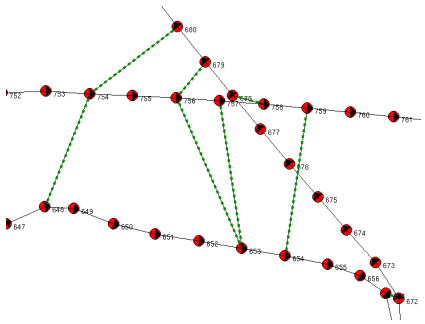

Fig. 1. The graph representation used. The figure shows frames (nodes) and relations (edges), both the odometry r_o and the visual relations r_v. Visual relations are indicated with dotted lines. Each frame a contains a reference to a set of features F_a extracted from an omni-directional image I_a, an odometry pose x_a^o, a covariance estimate of the odometry pose $C_{x_a^o}$, the estimated pose $\hat{x}_a$ and an estimate of its covariance $C_{\hat{x}_a}$. Fig. 2 shows images corresponding to the region represented by the graph in this figure.

Odometry relations r_o are created between successive frames. The relative pose μ_{r_o} is obtained directly from the odometry readings and the covariance C_{r_o} is estimated using the motion model suggested in [16] as

$$C_{r_o} = \begin{bmatrix} d^2\delta_{X_d}^2 + t^2\delta_{X_t}^2 & 0 & 0 \\ 0 & d^2\delta_{Y_d}^2 + t^2\delta_{Y_t}^2 & 0 \\ 0 & 0 & d^2\delta_{\theta_d}^2 + t^2\delta_{\theta_t}^2 \end{bmatrix} \quad (1)$$

where d and t are the total distance traveled and total angle rotated between two successive frames. The δ_X parameters relate to the forward motion, the δ_Y parameters to the side motion and the δ_θ parameters to the rotation of the robot. The six δ-parameters adjust the influence of the distance d and rotation t in the calculation of the covariance matrix. They were tuned manually once and then kept constant throughout the experiments.

C. Visual Similarity Relations

1) Similarity Measure: Given two images I_a and I_b, features are first extracted using the SIFT algorithm [4]. This results in two sets of features F_a and F_b for frame a and b. Each feature $F = [x, y], H$ comprises the pixel position $[x, y]$ and a histogram H containing the SIFT descriptor. The similarity measure $S_{a,b}$ is based on the number of features that match between F_a and F_b.

The feature matching algorithm calculates the Euclidean distance between each feature in image I_a and all the features in image I_b. A potential match is found if the smallest distance is smaller than 60% of the second smallest distance. This criterion was found empirically and was also used in [17]. It guarantees that interest point matches are substantially better than all other possible matches. We also do not allow features to be matched against more than one other feature. If a feature has more than one candidate match, the match which has the lowest Euclidean distance among the candidates is selected. Examples of matched features are shown in Fig. 2.

The matching step results in a set of feature pairs $P_{a,b}$ with a total number $M_{a,b}$ of matched pairs. Since the number of

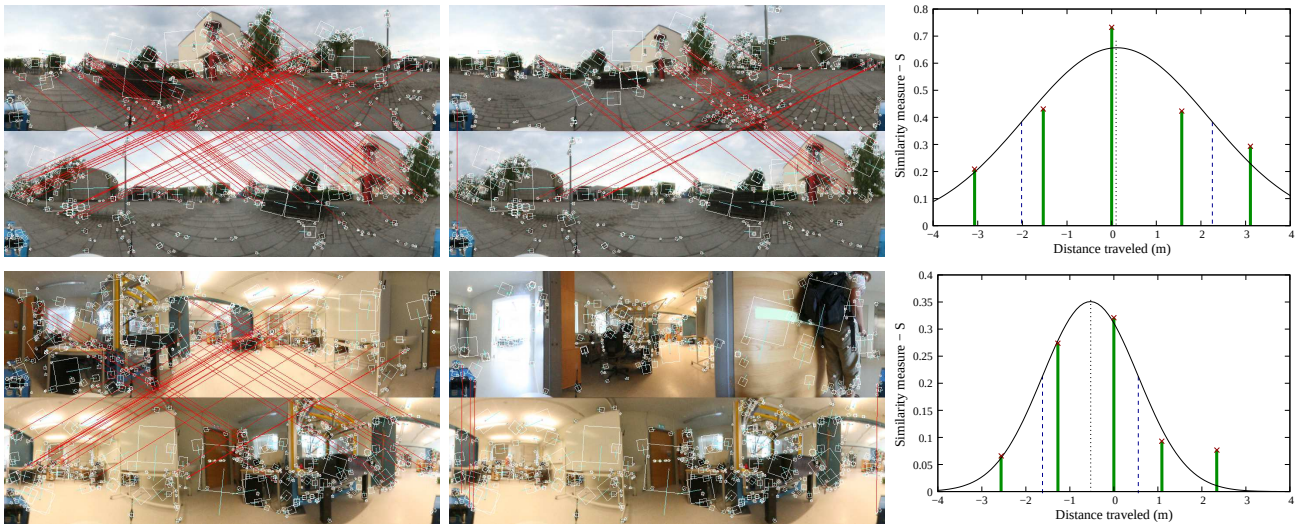

Fig. 2. Examples of loop closure detection outdoors (top) and indoors (bottom). In the outdoor example the distance to the extracted features is larger than in the indoor example. Left: feature matches at the peak of the similarity value, $S_{678,758} = 0.728$ (top) and $S_{7,360} = 0.322$ (bottom). Middle: feature matches two steps (equivalent to $\sim$3 meters distance) away, $S_{680,758} = 0.286$ (top) and $S_{9,360} = 0.076$ (bottom). The pose standard deviation $\sigma_{x^{r_v}} = \sigma_{y^{r_v}}$ was estimated as 2.06 m (top) and 1.09 m (bottom), respectively, and the mean d_μ as 0.199 m (top) and -0.534 m (bottom). Right: evolution of the similarity measure S against the distance travelled (obtained from odometry) together with the fitted Gaussian.

features varies heavily depending on the image content, the number of matches is normalized to $S_{a,b} \in [0, 1]$ as

$$S_{a,b} = \frac{M_{a,b}}{\frac{1}{2}(n_{F_a} + n_{F_b})} \tag{2}$$

where n_{F_a} and n_{F_b} are the number of features in F_a and F_b respectively. A high similarity measure indicates a perceptually similar position.

2) Estimation of the Relative Rotation and Variance: The relative rotation between two panoramic images I_a and I_b can be estimated directly from the horizontal displacement of the matched feature pairs $P_{a,b}$. If the flat floor assumption is violated this will be only an approximation. Here, the relative rotations θ_p for all matched pairs $p \in P_{a,b}$ are sorted into a 10 bin histogram and the relative rotation estimate $\mu_\theta^{r_v}$ is determined as the maximum of a parabola fitted to the largest bin and its left and right neighbour, see Fig. 3.

To evaluate the accuracy of relative rotation estimates θ_p, we collected panoramic images in an indoor laboratory environment and computed the relative orientation with respect to a reference image I_0. Panoramic images were recorded at a translational distance of 0.5, 1.0 and 2.0 meters to the reference image I_0. The ground truth rotation was obtained by manually measuring the displacement of corresponding pixels in areas along the displacement of the camera. The results in Table I demonstrate the good accuracy obtained. Even at a displacement of 2 meters the mean error is only 7.15 degrees.

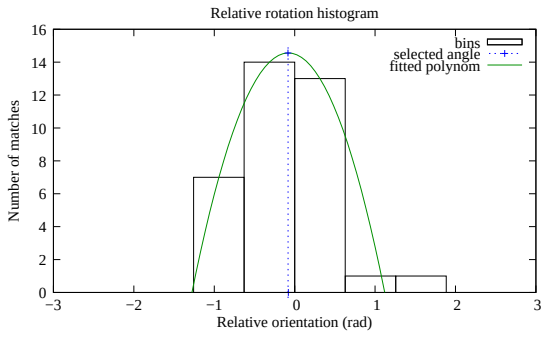

Fig. 3. Relative orientation histogram from two omnidirectional images taken 2 meters apart. The dotted line marks the relative orientation estimate $\mu_\theta^{r_v}$.

TABLE I
ERRORS OF RELATIVE ROTATION θ ESTIMATE IN RADIANS.

transl (m)	$error_\theta$	σ_{error_θ}
0.5	0.100	0.0630
1.0	0.104	0.0500
2.0	0.125	0.0903

The rotation variance $\sigma^2_{\theta^{r_v}}$ is estimated by the sum of squared differences between the estimate of the relative rotation $\mu_\theta^{r_v}$ and the relative rotation of the matched pairs $P_{a,b}$.

$$\sigma^2_{\theta^{r_v}} = \frac{1}{M_{a,b} - 1} \sum_{p \in P_{a,b}} (\mu_\theta^{r_v} - \theta_p)^2 \tag{3}$$

To increase the robustness towards outliers, a 10% Winsorized mean is applied. For the evalutated data this had only a minor effect on the results compared to using an un-truncated mean.

3) Estimation of Relative Position and Covariance: The Mini-SLAM approach does not attempt to determine the position of the detected features. Therefore, the relative position between two frames a and b cannot be determined very accurately. Instead we use only image similarity of the surrounding images to estimate $[\mu_x^{r_v}, \mu_y^{r_v}]$ as described below. It would be possible to estimate the relative position using multiple view geometry but this would introduce additional complexity that we want to avoid.

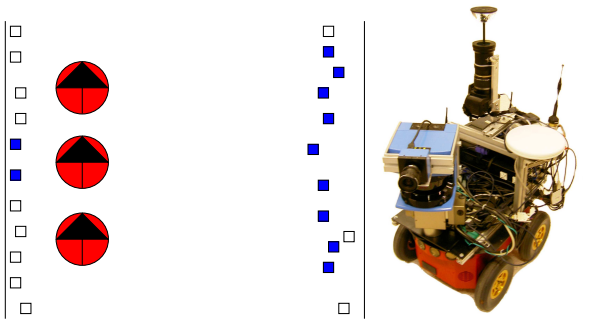

Fig. 4. Left: The physical distance to the features will influence the number of features that can be identified from different poses of the robot. The filled squares represent features that could be matched in all three robot poses while the unfilled squares represent the features for which correspondences could not be found from all poses. The left wall in the figure is closer to the robot. Thus, due to the faster change in appearance, the number of features of the left wall, which can be matched over successive images, tends to be less compared to the number of matched features of the right wall. Right: Outdoor robot used in this paper, equipped with a Canon EOS 350D camera and a panoramic lens from 0-360.com, which were used to collect the data, a DGPS unit to determine ground truth positions, and an LMS SICK scanner used for visualization and for obtaining ground truth.

Instead, geometric information is obtained from an estimate of the covariance of the relative position between a current frame b and a previously recorded frame a. This covariance estimate is computed using only the similarity measures S of frame b with a and the neighbouring frames of a.

The number of matched features between successive frames will vary depending on the physical distance to the features, see Figs. 2 and 4. Consider, for example, a robot located in an empty car park where the physical distance to the features is large and therefore the appearance of the environment does not change quickly if the robot is moved a certain distance. If, on the other hand, the robot is located in a narrow corridor where the physical distance to the extracted features is small, the number of feature matches in successive frames tends to be smaller if the robot was moved the same distance.

The covariance of the robot pose estimate [x,y]

$$C_{r_v} = \begin{bmatrix} \sigma^2_{x^{r_v}} & \sigma_{x^{r_v}}\sigma_{y^{r_v}} \\ \sigma_{x^{r_v}}\sigma_{y^{r_v}} & \sigma^2_{y^{r_v}} \end{bmatrix} \quad (4)$$

is computed based on how the similarity measure varies over the set $N(a)$, which contains frame a and its neighbouring frames. The analyzed sequence of similarity measures is indicated in the zoomed in visualization of a similarity matrix shown in Fig. 5. In order to avoid issues estimating the covariance orthogonal to the path of the robot if the robot was driven along a straight path, the covariance matrix is simplified by setting $\sigma^2_{x^{r_v}} = \sigma^2_{y^{r_v}}$ and $\sigma_{x^{r_v}}\sigma_{y^{r_v}} = 0$. The remaining covariance parameter is estimated by fitting a 1D Gaussian to the similarity measures $S_{N(a),b}$ and the distance travelled as obtained from odometry, see Fig. 6. Two parameters are determined from the nonlinear least squares fitting process: mean (d_μ) and variance ($\sigma^2_{[x,y]^r_v}$). The initial estimate of the relative position $[\mu^{r_v}_x, \mu^{r_v}_y]$ of a visual relation is calculated as

$$\mu^{r_v}_x = cos(\mu^{r_v}_\theta) d_\mu \quad (5)$$

$$\mu^{r_v}_y = sin(\mu^{r_v}_\theta) d_\mu, \quad (6)$$

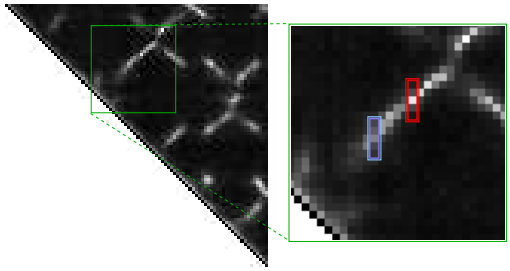

Fig. 5. Left: Full similarity matrix for the lab data set. Brighter entries indicate a higher similarity measure S. Right: Zoomed in image. The left area (enclosed in a blue frame) corresponds to a sequence of similarity measures that gives a larger position covariance than the right sequence (red frame).

where d_μ is the calculated mean of the fitted Gaussian and μ_θ the estimated relative orientation (Sec. II-C2).

In the experimental evaluation, the Gaussian was estimated using 5 consecutive frames. To evaluate whether the evolution of the similarity measure in the vicinity of a visual relation can be reasonably approximated by a Gaussian, the mean error between the 5 similarity measures and the fitted Gaussian was calculated for the outdoor/indoor data set (the data set is described in Sec. III-A). The results in Table II indicate that the Gaussian represents the evolution of the similarity in a reasonable way. Please note that frame b is recorded at a later time than frame a meaning that the covariance estimate $C^{a,b}_{r_v}$ can be calculated directly without any time lag.

4) Selecting Frames to Match: In order to speed up the algorithm and make it more robust to perceptual aliasing (the problem that different regions have similar appearance), only those frames are selected for matching that are likely to be

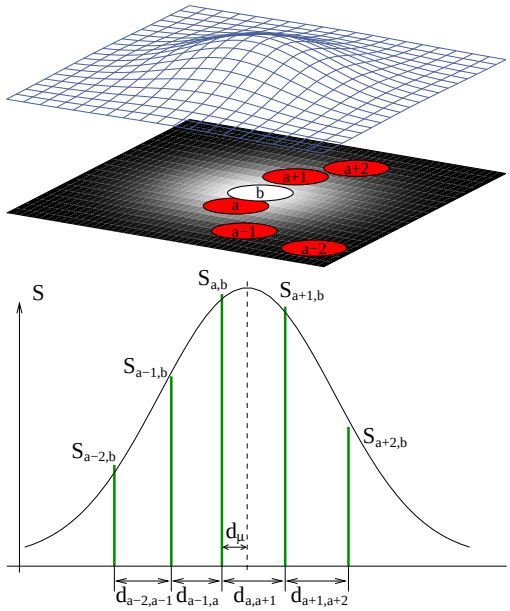

Fig. 6. Gaussian fitted to the distance travelled d (as obtained from odometry) and the similarity measures between frame b and the frames of the neighbourhood $N(a) = \{a-2, a-1, a, a+1, a+2\}$. From the similarity measures, both a relative pose estimate μ_{r^v} and a covariance estimate C_{r^v} are calculated between node a and node b. The orientation and orientation variance are not visualized in this figure.

TABLE II
STATISTICS OF THE ERROR ϵ BETWEEN THE GAUSSIAN FIT AND THE SIMILARITY MEASURES $S_{a-2,b}, ..., S_{a+2,b}$ FOR EACH NODE FOR WHICH THE FIT WAS PERFORMED IN THE OUTDOOR/INDOOR DATA SET.

node pair	ϵ	σ_ϵ
$< a-2, b >$	0.031	0.0441
$< a-1, b >$	0.029	0.0348
$< a, b >$	0.033	0.0601
$< a+1, b >$	0.026	0.0317
$< a+2, b >$	0.028	0.0388

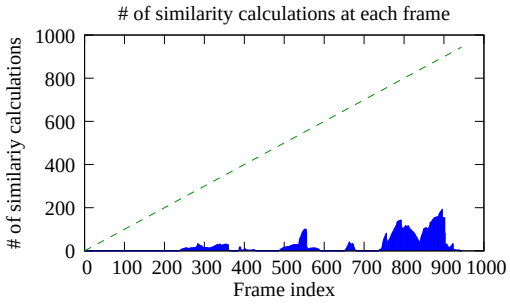

Fig. 7. Number of similarity calculations performed at each frame in the outdoor/indoor data set. The first frames were compared around frame 240, since up to then none of the previous frames were within the search area around the current pose estimate defined by the estimated pose covariance. The diagonal line indicates the linear increase for the case that the frames to match are not pre-selected.

located close to each other.

Consider the current frame b and a previously recorded frame a. If the similarity measure was to be calculated between b and all previously added frames, the number of frames to be compared would increase linearly, see Fig. 7. Instead, frames are only compared if the current frame b is within a search area around the pose estimate of frame a. The size of this search area is computed from the estimated pose covariance.

From the MLR algorithm (see Section II-A) we obtain the maximum likelihood estimate $\hat{x}_b$ for frame b. There is, however, no estimate of the corresponding covariance $C_{\hat{x}}$ that could be used to distinguish whether frame a is likely to be close enough to frame b so that it can be considered a candidate for a match, i.e. a frame for which the similarity measure $S_{a,b}$ should be calculated. So far, we have defined two types of covariances: the odometry covariance C_{r_o} and the visual relation covariance C_{r_v}. To obtain an overall estimate of the relative covariance between frame a and b we first consider the covariances of the odometry relations r_o between a and b and compute relative covariance $C_{x^o_{a,b}}$ as

$$C_{x^o_{a,b}} = \sum_{j \in (a,b-1)} \mathbf{R}_j C_{r_{o_j}} \mathbf{R}_j^T. \quad (7)$$

$\mathbf{R}_j$ is a rotation matrix, which is defined as

$$\mathbf{R}_j = \begin{pmatrix} cos(\hat{x}^\theta_{j+1} - \hat{x}^\theta_j) & -sin(\hat{x}^\theta_{j+1} - \hat{x}^\theta_j) & 0 \\ sin(\hat{x}^\theta_{j+1} - \hat{x}^\theta_j) & cos(\hat{x}^\theta_{j+1} - \hat{x}^\theta_j) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (8)$$

where $\hat{x}^\theta_j$ is the orientation estimated for frame j.

As long as no visual relation r_v has been added, either between a and b or any of the frames between a and b, the relative covariance $C_{\hat{x}_{a,b}}$ can be determined directly from the odometry covariance $C_{x^o_a}$ and $C_{x^o_b}$ as described above. However, when a visual relation $r_v^{a,b}$ between a and b is added, the covariance of the estimate $C_{\hat{x}_b}$ decreases. Using the covariance intersection method [18], the covariance for frame b is therefore updated as

$$C_{\hat{x}_b} = C_{\hat{x}_b} \oplus (C_{\hat{x}_a} + C_{r_v^{a,b}}), \quad (9)$$

where $\oplus$ is the covariance intersection operator. The covariance intersection method weighs the influence of both covariances C_a and C_b as

$$C_A \oplus C_B = [\omega C_A^{-1} + (1-\omega) C_B^{-1}]^{-1}. \quad (10)$$

The parameter $\omega \in [0, 1]$ is chosen so that the determinant of the resulting covariance is minimized [19].

The new covariance estimate is also used to update the frames between a and b by adding the odometry covariances $C_{x^o_{a..b}}$ in opposite order (i.e. simulate that the robot is moving backwards from frame b to a). The new covariance estimate for frame $j \in (a, b)$ is calculated as

$$C_{\hat{x}_j} = C_{\hat{x}_j} \oplus (C_{\hat{x}_b} + C_{x^o_{b,j}}). \quad (11)$$

5) Visual Relation Filtering: To avoid adding visual relations with low similarity, visual similarity relations $r_v^{a,b}$ between frame a and frame b are only added if the similarity measure exceeds a threshold t_{vs}: $S_{a,b} > t_{vs}$. In addition, similarity relations are only added if the similarity value $S_{a,b}$ has its peak at frame a (compared to the neighbouring frames $N(a)$). There is no limitation on the number of visual relations that can be added for each frame.

D. Fusing Multiple Data Sets

Fusion of multiple data sets recorded at different times is related to the problem of multi-robot mapping where each of the data sets is collected concurrently with a different robot. The motivation for multi-robot mapping is not only to reduce the time required to explore an environment but also to merge the different sensor readings in order to obtain a more accurate map. The problem addressed here is equivalent to "multi-robot SLAM with unknown initial poses" [20] because the relative poses between the data sets are not given. The exploration problem is not considered in this paper.

Only a minor modification of the standard method described above is necessary to address the problem of fusing multiple data sets. The absence of relative pose estimates between the data sets is compensated for by not limiting the search region for which similarity measures S are computed. This is implemented by incrementally adding data sets and setting the relative pose between consecutively added data sets initially to (0,0,0) with an infinite pose covariance. Such odometry relations between data sets appear as long, diagonal lines in Fig. 16 representing the transition between lab to $studarea$ and $studarea$ to $lab - studarea$.

III. Experimental Results

In this section, we present results from five different data sets with varying properties. An overview of all data sets is presented in Table III. All data sets were collected with our mobile robot Tjorven, see Fig. 4. The platform uses "skid-steering", which is prone to bad odometry. In the different data sets different wheel types (indoor / outdoor) were used. The robot's odometry was calibrated (for each wheel type) by first driving forward 5 meters to obtain a distance per encoder tick value, and second by completing one full revolution to determine the number of differential encoder ticks per angular rotation. Finally the drift parameter was adjusted so that the robot would drive forward in a straight line, i.e. to compensate for the slightly different size of the wheel pairs.

The omni-directional images were first converted to panoramic images with a resolution of 1000 x 289. When extracting SIFT features the initial doubling of the images was not performed, i.e. SIFT features from the first octave were ignored, simply to lower the amount of extracted features.

The results are presented both visually with maps obtained by superimposing laser range data using the poses estimated with Mini-SLAM and quantitatively by the mean squared error (MSE) from ground truth data. Since the corresponding pose pairs $< \hat{x}_i, x_i^{GT} >$ between the estimated pose $\hat{x}_i$ and the corresponding ground truth pose x_i^{GT} are known, the optimal rigid transformation between pose estimates and ground truth data can be determined directly. We applied the method suggested by Arun et al. [21].

To investigate the influence of the threshold t_{vs}, described in Section II-C5, the MSE was calculated for all data sets for which ground truth data were available. The result in Fig. 8 shows that the value of the threshold t_{vs} can be selected so that it is nearly optimal for all data sets and that there is a region in which minor changes of the t_{vs} do not strongly influence the accuracy of the map. Throughout the remainder of this section a constant threshold $t_{vs} = 0.2$ is used .

In order to give a better idea of the function of the Mini-SLAM algorithm, the number of visual relations per node depending on the threshold t_{vs} is shown in Fig. 9. The overview of all data sets presented in Table III also contains the number of similarity calculations performed and the evaluation run time on a Pentium 4 (2GHz) processor with 512 MB of RAM memory. This time does not include the time required for the similarity computation. Each similarity calculation (including relative rotation and variance estimation) took 0.30 seconds using a data set with an average of 522.3 features with standard deviation of 21.4. Please note, however, that the implementation used for feature matching in this paper was not optimised for computational efficiency.

TABLE III
For each data set: number of nodes $\#\hat{x}$, visual relations $\#r_v$, performed similarity calculations $\#S$, average number of extracted visual features μ_F per node with variance σ_F, evaluation run time T (excluding the similarity computation).

	$\#\hat{x}$	$\#r_v$	$\#S$	μ_F	σ_F	T (s)
outdoor / indoor	945	113	24784	497.5	170.0	66.4
multiple floor levels	409	198	13764	337.9	146.7	21.0
lab	86	60	443	571.5	39.6	3.6
$studarea$	134	31	827	426.6	51.1	9.4
$lab - studarea$	86	10	101	459.8	125.8	3.8

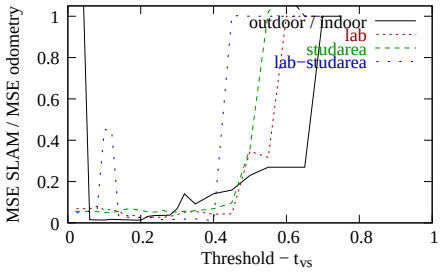

Fig. 8. The influence of the threshold parameter t_{vs} on the relative MSE.

A. Outdoor / indoor data set

A large set of 945 omni-directional images was collected over a total distance of 1.4 kilometers with height differences of up to 3 meters. The robot was driven manually and the data were collected in both indoor and outdoor areas over a period of 2 days (due to the limited capacity of the camera battery).

1) Comparison to ground truth obtained from DGPS: To evaluate the accuracy of the created map, the robot position was measured with differential GPS (DGPS) while collecting the omni-directional images. Thus, for every SLAM pose estimate there is a corresponding DGPS position $< \hat{x}_i, x_i^{DGPS} >$.

DGPS gives a smaller position error than GPS. However, since only the signal noise is corrected, the problem with multipath reflections still remains. DGPS is also only available if the radio link between the robot and the stationary GPS is functional. Thus, only a subset of pose pairs $< \hat{x}_i, x_i^{DGPS} >_{i=1..N}$ can be used for ground truth evaluation. DGPS measurements were considered only when at least five satellites were visible and the radio link to the stationary GPS was functional. The valid DGPS readings are indicated as light (blue) dots in Fig. 10. The total number of pairs used to calculate the MSE for the whole map was 377 compared to the total number of frames of 945. To measure the difference between the poses estimated with Mini-SLAM $\hat{x}$ and the DGPS positions x^{DGPS} (using UTM WGS84, which provides a metric coordinate system), the two data sets have to be aligned. Since the correspondence of the filtered pose pairs

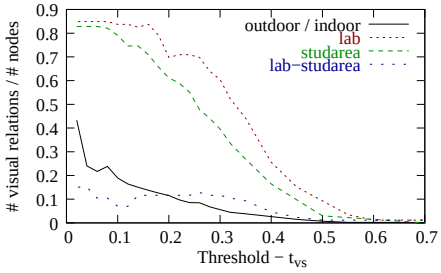

Fig. 9. The amount of visual nodes added to the graph depending on the threshold t_{vs}.

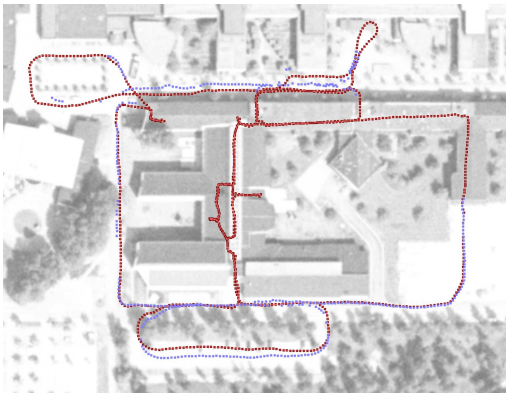

Fig. 10. DGPS data x^{DGPS} with aligned SLAM estimates $\hat{x}$ displayed on an aerial image of the area. The darker (red) squares show the Mini-SLAM pose estimates and the lighter (blue) squares show the DGPS poses for which the number of satellites was considered acceptable. The deviation seen at the bottom (the car park) is mainly caused by the fact that the car park is elevated compared to the rest of the environment.

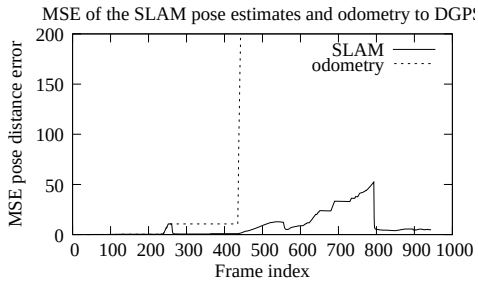

Fig. 11. Evolution of the MSE between the ground truth position obtained from DGPS readings x^{DGPS} and the Mini-SLAM estimate of the robot pose $\hat{x}$ as frames are added to the map. Drops in the MSE indicate that the consistency of the map has been increased. The final MSE of the raw odometry was 377.5 m^2.

is known, $< \hat{x}_i, x_i^{DGPS} >$, an optimal rigid alignment can be determined directly with the method by Arun et al. [21] as described above.

The mean square error (MSE) between x^{DGPS} and $\hat{x}$ for the data set shown in Fig. 10 is 4.89 meters. To see how it evolves over time when creating the map, the MSE was calculated from the new estimates $\hat{x}$ after each new frame was added. The result is shown in Fig. 11 and compared to the MSE obtained using only odometry to estimate the robot's position. Please note that the MSE was evaluated for each frame added. Therefore, when DGPS data are not available, the odometry MSE x^o will stay constant for these frames. This can be seen, for example, for the frames $250-440$ in Fig. 11. For the same frames, the MSE of the SLAM estimate $\hat{x}$ is not constant since new estimates are computed for each frame added and loop closing also occurs indoors or generally when no DGPS is available. The first visual relation r_v was added around frame 260. Until then, the error of the Mini-SLAM estimate $\hat{x}$ and the odometry MSE x^o were the same.

B. Multiple floor levels

This data set was collected inside a department building at Örebro University. It includes all (five) floor levels and connections between the floor levels by three elevators. The data contain loops in 2-d coordinates and also involving different floor levels. This data set consistst of 419 panoramic images and covers a path with a length of 618 meters. The geometrical layout differs for the different floors, see Fig. 13. No information about the floor level is used as an input to the system, hence the robot pose is still described using (x, y, θ).

1) Visualized results: There are no ground truth data available for this data set. It is possible, however, to get a visual impression of the accuracy of the results from Fig. 12. The figure shows occupancy grid maps obtained from laser scanner readings and raw odometry poses (left), or the Mini-SLAM pose estimates (right), respectively. All floors are drawn on top of each other without any alignment. To further illustrate the Mini-SLAM results, an occupancy map was also created separately for each floor from the laser scanner readings and Mini-SLAM pose estimates, see Fig. 13. Here, each pose was assigned to the corresponding floor level manually.

This experiment mainly illustrates the robustness of data association that is achieved using omni-directional vision data. The similarity matrix and a similarity access matrix for the "Multiple floor levels" data set are shown in Fig. 14.

C. Partly overlapping data

This data set consists of three separate indoor sets: lab (lab), student area ($studarea$) and a combination of both ($lab-studarea$), see Fig. 15. Similar to the data set described in Sec. III-B, omni-directional images, 2D laser range data and odometry were recorded. Ground truth poses x^{GT} were determined using the laser scanner and odometry together with the MLR approach as in [2].

1) Visualized results: Fig. 16 shows the final graph (left), a plot of laser scanner readings merged using poses from odometry (middle) and poses obtained with Mini-SLAM (right). Fig. 17 shows the similarity matrix and the similarity access matrix for the $lab-studarea$ data set.

2) Comparison to ground truth obtained from laser based SLAM: As described in Sec. II-D, fusion of multiple maps is motivated both by its need in multi-robot mapping and by the increased accuracy of the resulting maps. Instead of simply

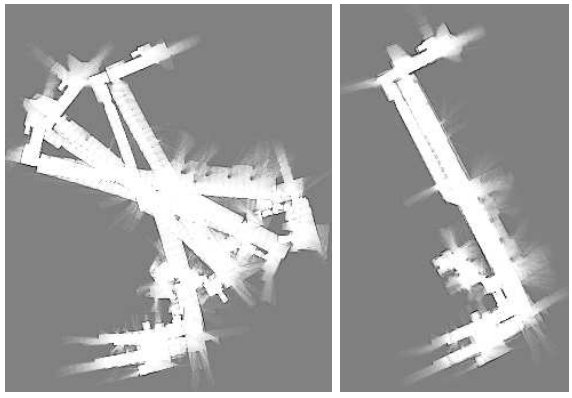

Fig. 12. Occupancy grid map of all five floors drawn on top of each other. Left: Gridmap created using pose information from raw odometry. Right: Using the estimated robot poses from Mini-SLAM.

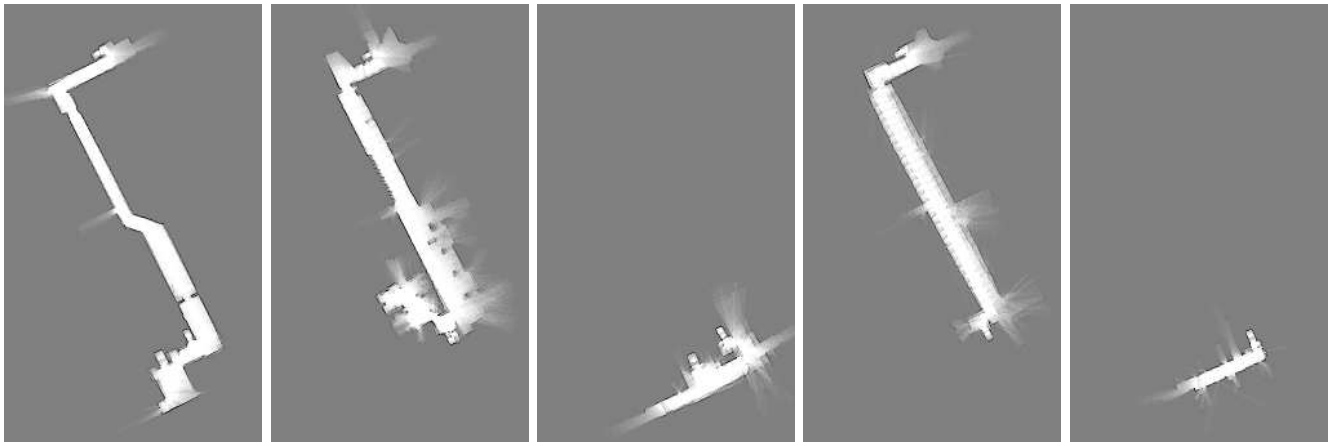

Fig. 13. Occupancy maps for floor levels 1-5, computed using laser scanner data at each estimated pose. The assignment of initial poses to floor levels was done manually and is only used to visualize these maps.

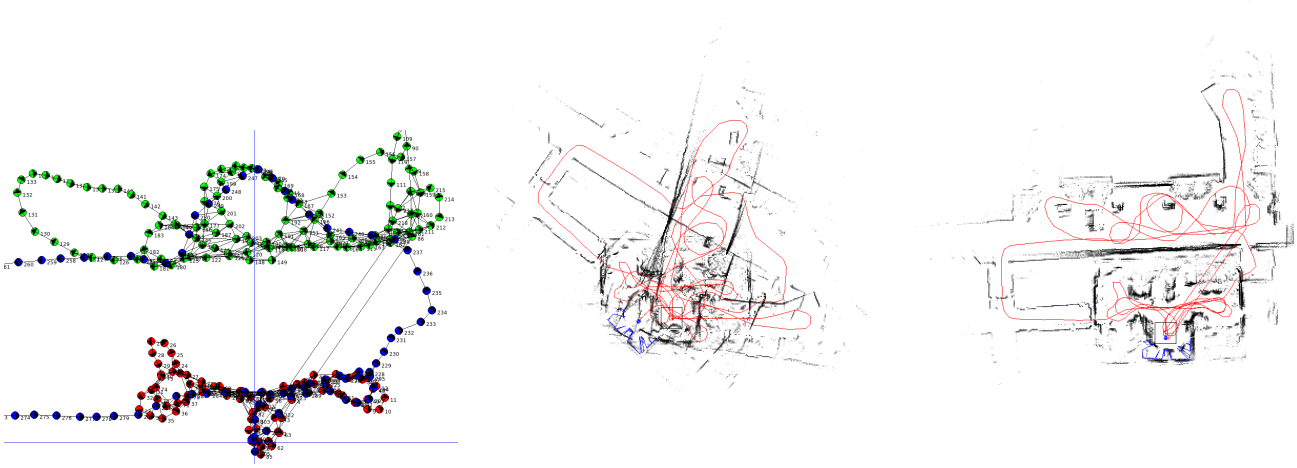

Fig. 16. Left: A part of the final MLR graph containing the three different data sets. Middle: Laser range scanning based map using the raw odometry. Right: Laser range scanning based map using the Mini-SLAM poses.

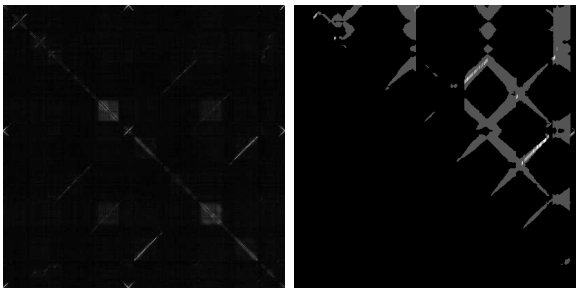

Fig. 14. Left: Pose similarity matrix for the "Multiple floor levels" data set. Right: Similarity access matrix showing which similarity measures were used in the Mini-SLAM computation. Brighter pixels were used more often.

adding the different maps onto each other, the fused maps also use additional information from the overlapping parts to improve the accuracy of the sub-maps. This is illustrated in Table IV which shows the MSE (again obtained by determining the rigid alignment between $\hat{x}$ and x^{GT}) before and after the fusion was performed. While the data sets lab and $studarea$ shows a negligible change in accuracy, $lab-studarea$ clearly demonstrate a large improvement.

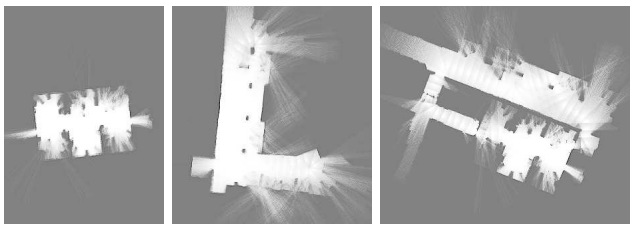

Fig. 15. Sub-maps for the partly overlapping data. Left: lab. Middle: $studarea$. Right: $lab-studarea$, overlapping both lab and $studarea$.

3) Robustness evaluation: The suggested method relies on incremental pose estimates (odometry) and a visual similarity

TABLE IV
MSE RESULTS BEFORE AND AFTER MERGING OF THE DATA SETS AND USING ODOMETRY ONLY.

	lab	$studarea$	$lab-studarea$
before fusion	0.002	0.029	0.036
after fusion	0.002	0.029	0.013
raw odometry	0.065	0.481	1.296

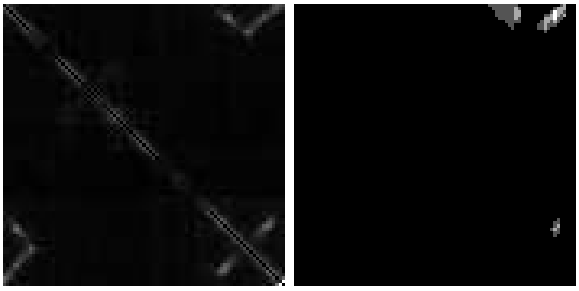

Fig. 17. Left: Pose similarity matrix for the $lab - studarea$ data set. Right: Similarity access matrix showing which similarity measures are used in the proposed method. Brighter pixels were used more often.

measure S. The robustness of the method is evaluated by corrupting these two inputs and evaluating the performance. For this evaluation, the $studarea$ data set is used and the tests were repeated 10 times.

In the first test, the similarity measures S were corrupted by adding a random value drawn from a Gaussian distribution $\mathcal{N}(0, \sigma)$ with varying standard deviation σ, see Table V. The amount of added noise has to be compared to the range of $[0, 1]$ in which the similarity measure S lies, see Eq. 2.

The robustness evaluation with respect to the similarity measure S shows that the system can handle additional noise to some extent, but incorrect visual relations will affect the accuracy of the final map. This illustrates that the proposed method, as many others, would have difficulties in perceptually similar locations in case the uncertainty of the pose estimates $C_{\hat{x}}$ is high.

In the second test, the odometry values were corrupted by adding additional noise to the incremental distance d and the orientation θ. The corrupted incremental distance d' is calculated as

$$d' = d + 0.1d\mathcal{N}(0, \sigma) + 0.2\theta\mathcal{N}(0, \sigma), \tag{12}$$

and the orientation θ' as

$$\theta' = \theta + 0.2d\mathcal{N}(0, \sigma) + \theta\mathcal{N}(0, \sigma). \tag{13}$$

Since the odometry pose estimates are computed incrementally the whole later trajectory is affected when adding noise at a particular time step.

The results of the robustness evaluation with the corrupted odometry are shown in Fig. 18 together with the MSE of the corrupted odometry. These results show that the system is robust to substantial odometry errors. A failure case is shown in Fig. 19.

TABLE V
MSE RESULTS ($mean$ AND $stddev$) AFTER ADDING A RANDOM VARIABLE DRAWN FROM $N(0, \sigma)$ TO EACH SIMILARITY MEASURE $S_{a,b}$.

σ	$mean$	$stddev$
0.02	0.03	0.004
0.05	0.03	0.011
0.10	0.11	0.074
0.20	0.94	0.992
0.40	1.35	1.304
0.80	1.49	1.240

IV. CONCLUSIONS AND FUTURE WORK

Mini-SLAM combines the principle of using similarity of panoramic images to close loops at the topological level with a graph relaxation method to obtain a metrically accurate map representation and with a novel method to determine the covariance for visual relations based on visual similarity of neighbouring poses. The proposed method uses visual similarity to compensate for the lack of range information about local image features, avoiding computationally expensive and less general methods such as tracking of individual image features.

Experimentally, the method scales well to the investigated environments. The experimental results are presented by visual means (as occupancy maps rendered from laser scans and poses determined by the Mini-SLAM algorithm) and by comparison with ground truth (obtained from DGPS outdoors or laser-based SLAM indoors). The results demonstrate that the Mini-SLAM method is able to produce topologically correct and geometrically accurate maps at low computational cost. A simple extension of the method was used to fuse multiple data sets so as to obtain improved accuracy. The method has also been used without any modifications to successfully map a building consisting of 5 floor levels.

Mini-SLAM generates a 2-d map based on 2-d input from odometry. It is worth noting that the "outdoor / indoor" data set includes variations of up to 3 meters in height. This indicates that the Mini-SLAM can cope with violations of the flat floor assumption to a certain extent. We expect a graceful degradation in map accuracy as the roughness of the terrain increases. The representation should still be useful for self-localization using 2-d odometry and image similarity, e.g., using the global localization method in [1], which in addition could be used to improve the robustness towards perceptual aliasing when fusing multiple data sets. In extreme cases, of course, it is possible that the method would create inconsistent maps, and a 3-d representation should be considered.

The bottleneck of the current implementation in terms of computation time is the calculation of image similarity, which involves the comparison of many local features. The suggested approach, however, is not limited to the particular measure of image similarity used in this work. There are

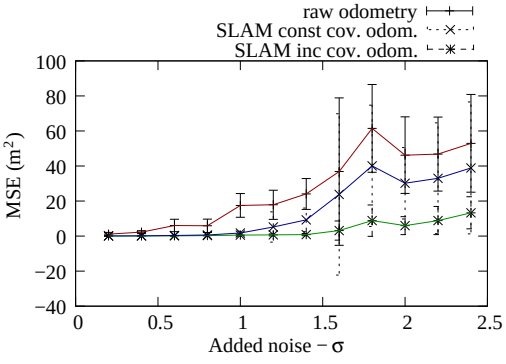

Fig. 18. MSE results ($mean$ and $stddev$) for x (odometry) and $\hat{x}$ (estimated poses) after corrupting the odometry by adding random values drawn from $N(0, \sigma)$. The plot also shows the MSE when the odometry covariance is increased with the added noise.

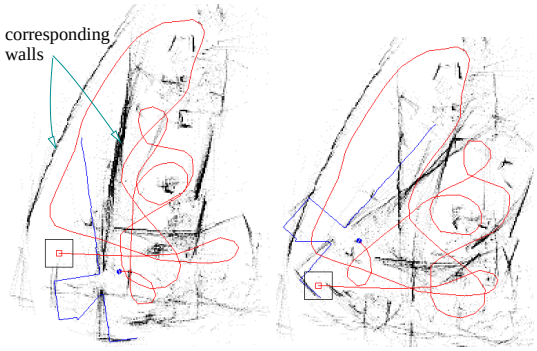

Fig. 19. A failure case where the corrupted odometry error became too large resulting in a corrupted map. Left: SLAM map. Right: raw odometry.

many possibilities to increase the computation speed either by using alternative similarity measures that are faster to compute while still being distinctive enough, or by optimizing the implementation, for example, by executing image comparisons on a graphics processing unit (GPU) [22].

Further plans for future work include an investigation of the possibility of using a standard camera instead of an omni-directional camera, and incorporation of vision-based odometry to realise a completely vision-based system.

REFERENCES

[1] H. Andreasson, A. Treptow, and T. Duckett, "Localization for mobile robots using panoramic vision, local features and particle filter," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2005, pp. 3348–3353.

[2] U. Frese, P. Larsson, and T. Duckett, "A multilevel relaxation algorithm for simultaneous localisation and mapping," *IEEE Transactions on Robotics*, vol. 21, no. 2, pp. 196–207, April 2005.

[3] H. Andreasson, T. Duckett, and A. Lilienthal, "Mini-SLAM: Minimalistic visual SLAM in large-scale environments based on a new interpretation of image similarity," in *IEEE International Conference on Robotics and Automation (ICRA 2007)*, Rome, Italy, 2007.

[4] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Computer Vision ICCV, Corfu*, 1999, pp. 1150–1157.

[5] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *International Journal of Robotics Research*, vol. 21, no. 8, pp. 735–758, 2002.

[6] T. Barfoot, "Online visual motion estimation using FastSLAM with SIFT features," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2005, pp. 579–585.

[7] P. Jensfelt, D. Kragic, J. Folkesson, and M. Björkman, "A framework for vision based bearing only 3D SLAM," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2006, pp. 1944–1950.

[8] N. Karlsson, E. D. Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. E. Munich, "The vSLAM algorithm for robust localization and mapping," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2005, pp. 24–29.

[9] P. Elinas, R. Sim, and J. Little, "σSLAM: Stereo vision SLAM using the Rao-Blackwellised particle filter and a novel mixture proposal distribution," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2006, pp. 1564–1570.

[10] J. Sez and F. Escolano, "6dof entropy minimization slam," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2006, pp. 1548–1555.

[11] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV'03)*, 2003, pp. 1403–1410.

[12] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous System*, vol. 54, no. 9, pp. 740–749, September 2006.

[13] P. M. Newman, D. M. Cole, and K. L. Ho, "Outdoor SLAM using visual appearance and laser ranging," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2006, pp. 1180–1187.

[14] G. Grisetti, D. Lordi Rizzini, C. Stachniss, E. Olson, and W. Burgard, "Online constraint network optimization for efficient maximum likelihood map learning," Pasadena, CA, USA, 2008.

[15] E. Olson, J. Leonard, and S. Teller, "Fast iterative optimization of pose graphs with poor initial estimates," 2006, pp. 2262–2269.

[16] A. I. Eliazar and R. Parr, "Learning probabilistic motion models for mobile robots," in *Proc. of the twenty-first Int. Conf. on Machine learning (ICML)*, 2004, p. 32.

[17] J. Gonzalez-Barbosa and S. Lacroix, "Rover localization in natural environments by indexing panoramic images," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*. IEEE, 2002, pp. 1365–1370.

[18] J. Uhlmann, "Dynamic map building and localization for autonomous vehicles," Ph.D. dissertation, University of Oxford, 1995.

[19] S. J. Julier and J. K. Uhlmann, "Using covariance intersection for slam," *Robotics and Autonomous Systems*, vol. 55, no. 1, pp. 3–20, 2007.

[20] A. Howard, "Multi-robot simultaneous localization and mapping using particle filters," in *Proceedings of Robotics: Science and Systems*, Cambridge, USA, June 2005.

[21] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, 1987.

[22] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc, "GPU-based video feature tracking and matching," in *Workshop on Edge Computing Using New Commodity Architectures (EDGE 2006)*, Chapel Hill, 2006.

Henrik Andreasson is a Ph.D. student at Centre for Applied Autonomous Sensor System, Örebro University, Sweden. He received his Master degree in Mechatronics from Royal Institute of Technology, Sweden, in 2001. His research interests include mobile robotics, computer vision, and machine learning.

Tom Duckett is a Reader at the Department of Computing and Informatics, University of Lincoln, where he is also Director of the Centre for Vision and Robotics Research. He was formerly a docent (Associate Professor) at Örebro University, where he was also leader of the Learning Systems Laboratory within the Centre for Applied Autonomous Sensor Systems. He obtained his Ph.D. from Manchester University, M.Sc. with distinction from Heriot-Watt University and B.Sc. (Hons.) from Warwick University, and has also studied at Karlsruhe and Bremen Universities. His research interests include mobile robotics, navigation, machine learning, AI, computer vision, and sensor fusion for perception-based control of autonomous systems.

Achim Lilienthal is a docent (associate professor) at the AASS Research Center, Örebro, Sweden where he is leading the Learning Systems Lab. He obtained his Ph.D. in computer science from Tübingen University, Germany and his M.Sc. and B.Sc. in Physics from the University of Konstanz, Germany. The Ph.D. thesis addresses gas distribution mapping and gas source localisation with a mobile robot. The M.Sc. thesis is concerned with an investigation of the structure of $(C_{60})_n^+$ clusters using gas phase ion chromatography. His main research interests are mobile robot olfaction, robot vision, robotic map learning and safe navigation systems.